Model Assisted Statistics and Applications 9 (2014) 25–35
DOI 10.3233/MAS-130276
IOS Press

An estimation of a sensitive attribute by two stage stratified randomized response model

Ki-Hak Hong[a], Gi-Sung Lee[b], Chang-Kyoon Son[c] and Jong-Min Kim[d,*]
[a]*Department of Computer Science, Dongshin University, Naju, Chonnam, Korea*
[b]*Department of Children Welfare, Woosuk University, Wanju-gun, Jeonbuk, Korea*
[c]*Department of Statistics and Information Science, Dongguk University, Gyeongju, Gyeongbuk, Korea*
[d]*Statistics Discipline, Division of Science and Mathematics, University of Minnesota – Morris, Morris, MN, USA*

Abstract. We deal with the estimation of sensitive attribute of the population which is composed of a number of strata by applying stratified sampling to Abdelfatah et al.'s model [1]. We estimate the sensitive parameter in the case of knowing the size of stratum, and check the effect of the proportional allocation method and the optimum allocation method. We extend it to the case of not knowing the size of stratum, and estimate the sensitive parameter by applying stratified double sampling to Abdelfatah et al.'s model [1]. Finally, we compare the efficiency of our suggested estimator to the existing Abdelfatah et al.'s estimator. A practical problem with the use of optimum allocation has been pointed out. Thus, in practice, the use of either proportional allocation or equal allocation has been suggested while estimating proportion of a sensitive attribute using stratified randomized response sampling.

Keywords: Randomized response model, two-stage model, sensitive attribute, stratified sampling, stratified double sampling

1. Introduction

Warner [13] first suggested an ingenious survey model called the randomized response model (RRM) to procure sensitive information from respondents without disturbing their privacy by using a randomizing device composed of two questions; one was sensitive and the other was non-sensitive:

Do you have a sensitive attribute A? (with probability P_0)

Do you have a nonsensitive attribute $\bar{A}$? (with probability $1 - P_0$)

To estimate the population proportion π belonging to the sensitive group (A), a simple random sample with replacement (SRSWR) of n persons is drawn from the population, and the respondents are required to answer "Yes" or "No" according to the results of randomizing device not revealed to the interviewer.

The maximum likelihood estimator of π and its variance are:

$$\begin{aligned} \hat{\pi}_w &= \frac{n'/n - (1 - P_0)}{2P_0 - 1}, P_0 \neq 1/2, \\ V(\hat{\pi}_w) &= \frac{\pi(1-\pi)}{n} + \frac{P_0(1-P_0)}{n(2P_0-1)^2}, \end{aligned} \tag{1}$$

where n' is the number of "Yes" answers obtained from the n respondents.

*Corresponding author: Ki-Hak Hong, Statistics Discipline, Division of Science and Mathematics, University of Minnesota – Morris, Morris, MN 56267, USA. E-mail: jongmink@morris.umn.edu.

ISSN 1574-1699/14/$27.50

Table 1
Classification of responses from Deck(1) and Deck(2)

Responses from Deck(1)	Responses from Deck(2)	
	Yes	No
Yes	n_{11}	n_{10}
No	n_{01}	n_{00}

Mangat-Singh [8] developed a two stage randomized response model to increase the efficiency of Warner's model. In this method, each interviewee in the SRSWR of n respondents is provided with two randomizing devices:

[*Randomizing device* R_1]

Statement 1: Do you have a sensitive attribute A? (with probability T_0)

Statement 2: Go to randomizing device R_2 (with probability $1 - T_0$)

[*Randomizing device* R_2]

Statement 1: Do you have a sensitive attribute A? (with probability P_0)

Statement 2: Do you have a nonsensitive attribute $\bar{A}$? (with probability $1 - P_0$)

The respondent is required to answer "Yes" or "No" according to the statement and the actual status he/she possesses. The maximum likelihood estimator of π and its variance are:

$$\begin{aligned} \hat{\pi}_{ms} &= \frac{n'/n - (1 - T_0)(1 - P_0)}{2P_0 - 1 + 2\,T_0(1 - P_0)}, \\ V(\hat{\pi}_{ms}) &= \frac{\pi(1 - \pi)}{n} + \frac{(1 - T_0)(1 - P_0)[1 - (1 - T_0)(1 - P_0)]}{n[2P_0 - 1 + 2\,T_0(1 - P_0)]^2}, \end{aligned} \tag{2}$$

where n' is the number of "Yes" answers obtained from the n respondents.

Mangat-Singh showed that their model was more efficient than Warner's model under the condition $T_0 > (1 - -2P_0)/(1 - P_0)$. Mangat [7] developed a randomized response model which reduced the use of the randomizing device from two to one. Each of n respondents selected with a SRSWR is instructed to say "Yes" if he/she has the attribute A. If the respondent doesn't have the attribute A, then he/she is required to answer according to the question selected from the Warner's randomization device.

The maximum likelihood estimator of π and its variance are:

$$\begin{aligned} \hat{\pi}_m &= \frac{n'/n - (1 - P_0)}{P_0}, \\ V(\hat{\pi}_m) &= \frac{\pi(1 - \pi)}{n} + \frac{(1 - \pi)(1 - P_0)}{nP_0}, \end{aligned} \tag{3}$$

where n' is the number of "Yes" answers obtained from the n respondents.

Odumade and Singh [9] suggested the use of two decks of cards in a randomized response model where each of the decks included the two statements used in Warner's [13] model. Each respondent in a SRSWR of n respondents is provided with two decks of cards. Deck(1) includes the two questions, (a) Do you have a sensitive attribute A? (b) Do you have a nonsensitive attribute $\bar{A}$? with probabilities P and $1 - P$ respectively. Deck(2) includes the two questions as in Deck(1) with probabilities T and $1 - T$ respectively. Each respondent is requested to draw two cards simultaneously, one by one from each deck of cards, and read the questions in order and answer "Yes" or "No" according to them. The responses from the n respondents can be classified into the 2×2 contingency table as shown in Table 1.

An unbiased estimator of the population proportion π and its variance is given by:

$$\begin{aligned} \hat{\pi}_{os} &= \frac{1}{2} + \frac{(P + T - 1)(n_{11}/n - n_{00}/n) + (P - T)(n_{10}/n - n_{01}/n)}{2[(P + T - 1)^2 + (P - T)^2]}, P \neq \frac{1}{2}, T \neq \frac{1}{2}, \\ V(\hat{\pi}_{os}) &= \frac{(P + T - 1)^2[PT + (1 - P)(1 - T)] + (P - T)^2[T(1 - P) + P(1 - T)]}{4n[(P + T - 1)^2 + (P - T)^2]^2} - \frac{(2n - 1)^2}{4n}, \end{aligned}$$

$$P \neq \frac{1}{2}, T \neq \frac{1}{2}. \tag{4}$$

Abdelfatah et al. [1] suggested a modified Odumade and Singh [9] model by using Mangat-Singh's [8] procedure instead of Warner's procedure in each stage. We deal with the estimation of a sensitive attribute of the population, which is composed of a number of strata, by adapting the procedure of Abdelfatah et al.'s model [1]. We estimate the sensitive parameter in the case of knowing the size of stratum, and check the effect of the proportional allocation method and the optimum allocation method. We extend it to the case of not knowing the size of stratum and estimate the sensitive parameter by applying stratified double sampling to Abdelfatah et al.'s model [1]. Finally, we compare the efficiency of our suggested method with the existing Abdelfatah et al.'s model [1].

2. A stratified estimation of a sensitive attribute by two stage randomized response model

When the population is composed of a number of strata, we deal with the estimation of sensitive attribute of the population by adapting the two stage randomized response model suggested by Abdelfatah et al. [1]. We estimate the sensitive parameter in the case of knowing the size of stratum and check the effect of the proportional allocation method and the optimum allocation method.

Let the population be composed of a number of mutually disjoint L strata of $N_h (h = 1, 2, \ldots, L)$. Each of $n_h (n = \sum_{h=1}^{L} n_h)$ respondents selected by a SRSWR from $N_h (h = 1, 2, \ldots, L)$ is requested to draw two cards simultaneously, one card from each deck of cards, and read the statements in order. The respondent is requested to draw a card from Deck(3) only if directed by the outcome of Deck(1), and he/she is also requested to draw a card from Deck(4) only if directed by the outcome of Deck(2). Deck(3) and Deck(4) are exactly the same decks used by Odumade and Singh [9]. The respondent first matches his/her actual status with the question written on the card drawn from Deck(1) or Deck(3), and then he/she matches his/her actual status with the question written on the card drawn from Deck(2) or Deck(4). The whole procedure is done completely by the respondent, away from the interviewer.

Since the response (Yes, Yes) from stratum h can be answered from any respondent, regardless of having sensitive attribute A_h, the interviewer can't know the interviewee's real status.

The probability of getting (Yes, Yes) response from stratum h, θ_{h11} is given by

$$\begin{aligned}
\theta_{h11} &= P(\textit{Yes}, \textit{Yes}) \\
&= W_h Q_h \pi_h + W_h (1 - Q_h) T_h \pi_h + (1 - W_h) P_h Q_h \pi_h + (1 - W_h) P_h (1 - Q_h) T_h \pi_h \\
&\quad + (1 - W_h)(1 - P_h)(1 - Q_h)(1 - T_h)(1 - \pi_h) \\
&= [(1 - W_h) P_h + (1 - Q_h) T_h + Q_h + W_h - 1] \pi_h + (1 - W_h)(1 - P_h)(1 - Q_h)(1 - T_h)
\end{aligned}$$

In the same way, the probabilities, θ_{h10}, θ_{h01}, θ_{h00} are given by

$$\begin{aligned}
\theta_{h10} &= P(\textit{Yes}, \textit{No}) \\
&= [W_h - Q_h + P_h (1 - W_h) - T_h (1 - Q_h)] \pi_h + (1 - W_h)(1 - P_h)[Q_h + (1 - Q_h) T_h],
\end{aligned}$$

$$\begin{aligned}
\theta_{h01} &= P(\textit{No}, \textit{Yes}) \\
&= [Q_h - W_h + T_h (1 - Q_h) - P_h (1 - W_h)] \pi_h + (1 - Q_h)(1 - T_h)[W_h + (1 - W_h) P_h],
\end{aligned}$$

and

$$\begin{aligned}
\theta_{h00} &= P(No, No) \\
&= [1 - W_h - Q_h - P_h (1 - W_h) - T_h (1 - Q_h)] \pi_h \\
&\quad + W_h Q_h + W_h (1 - Q_h) T_h + (1 - W_h) P_h Q_h + (1 - W_h) P_h (1 - Q_h) T_h.
\end{aligned}$$

Table 2
Classification of the responses from the four decks of cards in stratum h

Responses from decks (1 or 3)	Responses from decks (2 or 4)		Total
	Yes	No	
Yes	n_{h11}	n_{h10}	$n_{h1\cdot}$
No	n_{h01}	n_{h00}	$n_{h0\cdot}$
Total	$n_{h\cdot 1}$	$n_{h\cdot 0}$	n_h

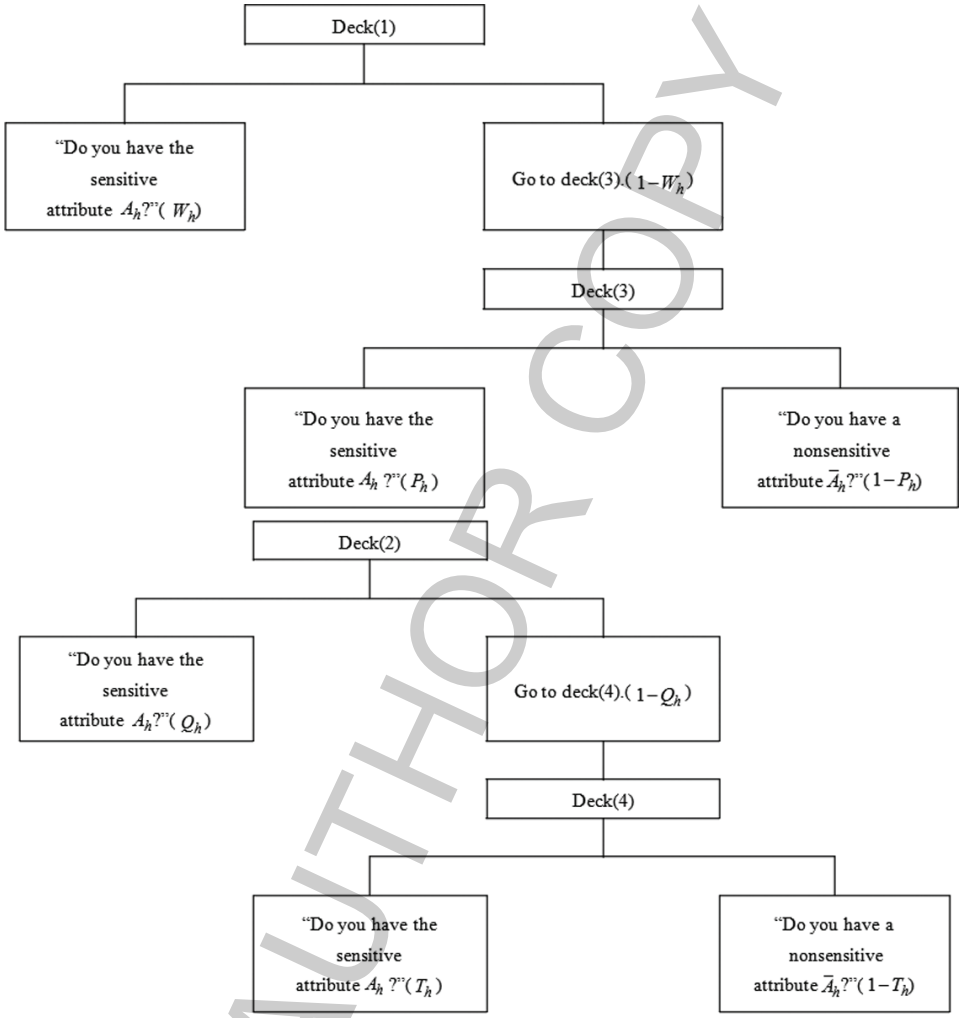

Fig. 1. The randomization device of stratum h.

The responses from the n_h respondents of stratum h can be classified into a 2×2 contingency table as shown in Table 2. In order to estimate the unknown population proportion π_h of the respondents belonging to sensitive group A_h in stratum h, let n_{h11}/n_h, n_{h10}/n_h, n_{h01}/n_h, n_{h00}/n_h be the observed proportions of (Yes, Yes), (Yes, No), (No, Yes) and (No, No) responses as unbiased estimators for $\theta_{h11}, \theta_{h10}, \theta_{h01}$ and θ_{h00} respectively, where $\sum_{h=1}^{L}\sum_{i=0}^{1}\sum_{j=0}^{1}\theta_{hij} = 1$.

We define the local squared distance between the observed and the true proportions in each stratum h as:

$$D_h = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}(\theta_{hij} - n_{hij}/n_h)^2,$$

where

$$
\begin{aligned}
D_h = \frac{1}{2}\Bigg\{ & \left[(1-W_h)P_h+(1-Q_h)T_h+Q_h+W_h-1\right]\pi_h+(1-W_h)(1-P_h)(1-Q_h)(1-T_h)-\frac{n_{h11}}{n_h}\Bigg\}^2 \\
& + \frac{1}{2}\Bigg\{ \left[W_h-Q_h+P_h(1-W_h)-T_h(1-Q_h)\right]\pi_h+(1-W_h)(1-P_h)\left[(Q_h+(1-Q_h)T_h\right]-\frac{n_{h10}}{n_h}\Bigg\}^2 \\
& + \frac{1}{2}\Bigg\{ \left[Q_h-W_h+T_h(1-Q_h)-P_h(1-W_h)\right]\pi_h+(1-Q_h)(1-T_h)\left[W_h+(1-W_h)P_h\right]-\frac{n_{h01}}{n_h}\Bigg\}^2 \\
& + \frac{1}{2}\Big\{ \left[1-W_h-Q_h-P_h(1-W_h)-T_h(1-Q_h)\right]\pi_h+W_hQ_h+W_h(1-Q_h)T_h+(1-W_h)P_hQ_h \\
& +(1-W_h)P_h(1-Q_h)T_h-\frac{n_{h00}}{n_h}\Bigg\}^2 .
\end{aligned}
$$

To obtain the π_h that minimizes the local squared distance D_h, we have

$$
\begin{aligned}
\frac{\partial D_h}{\partial \pi_h} = & \left[(1-W_h)P_h+(1-Q_h)T_h+Q_h+W_h-1\right]^2\pi_h-\frac{n_{h11}}{n_h}\left[(1-W_h)P_h+(1-Q_h)T_h+Q_h+W_h-1\right] \\
& + \left[(1-W_h)P_h+(1-Q_h)T_h+Q_h+W_h-1\right](1-W_h)(1-P_h)(1-Q_h)(1-T_h) \\
& + \left[W_h-Q_h+P_h(1-W_h)-T_h(1-Q_h)\right]^2\pi_h-\frac{n_{h10}}{n_h}\left[W_h-Q_h+P_h(1-W_h)-T_h(1-Q_h)\right] \\
& + \left[W_h-Q_h+P_h(1-W_h)-T_h(1-Q_h)\right](1-W_h)(1-P_h)\left[Q_h+(1-Q_h)T_h\right] \\
& + \left[Q_h-W_h+T_h(1-Q_h)-P_h(1-W_h)\right]^2\pi_h-\frac{n_{h01}}{n_h}\left[Q_h-W_h+T_h(1-Q_h)-P_h(1-W_h)\right] \\
& + \left[Q_h-W_h+T_h(1-Q_h)-P_h(1-W_h)\right](1-Q_h)(1-T_h)\left[W_h+(1-W_h)P_h\right] \\
& + \left[1-W_h-Q_h-P_h(1-W_h)-T_h(1-Q_h)\right]^2\pi_h-\frac{n_{h00}}{n_h}\left[1-W_h-Q_h-P_h(1-W_h)-T_h(1-Q_h)\right] \\
& + \left[1-W_h-Q_h-P_h(1-W_h)-T_h(1-Q_h)\right]\left[W_hQ_h+W_h(1-Q_h)T_h\right. \\
& \left.+(1-W_h)P_hQ_h+(1-W_h)P_h(1-Q_h)T_h\right],
\end{aligned}
$$

and setting $\frac{\partial D_h}{\partial \pi_h}=0$, we obtain the following estimator $\hat{\pi}_h$ of the population proportion π_h in stratum h:

$$
\hat{\pi}_h = \frac{1}{2}+\frac{(n_{h11}/n_h-n_{h00}/n_h)B_h+(n_{h10}/n_h-n_{h01}/n_h)C_h}{2(B_h^2+C_h^2)},
$$

where $B_h=(1-W_h)P_h+(1-Q_h)T_h+W_h+Q_h-1, C_h=W_h-Q_h+(1-W_h)P_h-(1-Q_h)T_h$.

Then, the overall estimator $\hat{\pi}$ of the population proportion π is obtained by

$$
\hat{\pi}=\sum_{h=1}^{L} Z_h\left[\frac{1}{2}+\frac{(n_{h11}/n_h-n_{h00}/n_h)B_h+(n_{h10}/n_h-n_{h01}/n_h)C_h}{2(B_h^2+C_h^2)}\right], Z_h=\frac{N_h}{N}. \tag{5}
$$

Theorem 1. The estimator $\hat{\pi}$ is an unbiased estimator of the population proportion π.

Proof It follows from the fact that $E(n_{hij}/n)=\theta_{hij}$, $h=1,2,\ldots,L; i=0,1; j=0,1$.

$$
\begin{aligned}
E(\hat{\pi}) &= E\left[\sum_{h=1}^{L} Z_h\left(\frac{1}{2}+\frac{(n_{h11}/n_h-n_{h00}/n_h)B_h+(n_{h10}/n_h-n_{h01}/n_h)C_h}{2(B_h^2+C_h^2)}\right)\right] \\
&= \sum_{h=1}^{L} Z_h\left[\frac{1}{2}+\frac{(\theta_{h11}-\theta_{h00})B_h+(\theta_{h10}-\theta_{h01})C_h}{2(B_h^2+C_h^2)}\right]=\sum_{h=1}^{L} Z_h\pi_h=\pi
\end{aligned}
$$

Theorem 2. The variance of the estimator $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \sum_{h=1}^{L} Z_h^2 \left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4n_h(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4n_h} \right], \tag{6}$$

where

$$\begin{aligned}
B_h &= (1 - W_h)P_h + (1 - Q_h)T_h + W_h + Q_h - 1, \\
C_h &= W_h - Q_h + (1 - W_h)P_h - (1 - Q_h)T_h, \\
E_h &= W_h Q_h + W_h(1 - Q_h)T_h + (1 - W_h)P_h Q_h + (1 - W_h)P_h(1 - Q_h)T_h, \\
F_h &= (1 - W_h)(1 - P_h)(1 - Q_h)(1 - T_h), \\
G_h &= (1 - Q_h)(1 - T_h)[W_h + (1 - W_h)P_h], \\
H_h &= (1 - W_h)(1 - P_h)[Q_h + (1 - Q_h)T_h].
\end{aligned}$$

Proof

$$\begin{aligned}
V(\hat{\pi}) &= V\left[\sum_{h=1}^{L} Z_h \left(\frac{1}{2} + \frac{(n_{h11}/n_h - n_{h00}/n_h)B_h + (n_{h10}/n_h - n_{h01}/n_h)C_h}{2(B_h^2 + C_h^2)} \right) \right] \\
&= \sum_{h=1}^{L} \frac{Z_h^2}{4(B_h^2 + C_h^2)^2} \left[V\left(\frac{n_{h11}}{n_h} - \frac{n_{h00}}{n_h} \right) B_h^2 + V\left(\frac{n_{h10}}{n_h} - \frac{n_{h01}}{n_h} \right) C_h^2 \right. \\
&\quad \left. + 2B_h C_h Cov\left(\frac{n_{h11} - n_{h00}}{n_h}, \frac{n_{h10} - n_{h01}}{n_h} \right) \right]
\end{aligned}$$

Using the following results from the multinomial distribution, we can prove <Theorem 2>.

$$\begin{aligned}
&V(n_{h11}/n_h) = \theta_{h11}(1 - \theta_{h11})/n_h, V(n_{h10}/n_h) = \theta_{h10}(1 - \theta_{h10})/n_h, \\
&V(n_{h01}/n_h) = \theta_{h01}(1 - \theta_{h01})/n_h, V(n_{h00}/n_h) = \theta_{h00}(1 - \theta_{h00})/n_h, \\
&Cov(n_{h11}/n_h, n_{h10}/n_h) = -\theta_{h11}\theta_{h10}/n_h, Cov(n_{h10}/n_h, n_{h01}/n_h) = -\theta_{h10}\theta_{h01}/n_h, \\
&Cov(n_{h11}/n_h, n_{h01}/n_h) = -\theta_{h11}\theta_{h01}/n_h, Cov(n_{h10}/n_h, n_{h00}/n_h) = -\theta_{h10}\theta_{h00}/n_h, \\
&Cov(n_{h01}/n_h, n_{h00}/n_h) = -\theta_{h01}\theta_{h00}/n_h, Cov(n_{h11}/n_h, n_{h00}/n_h) = -\theta_{h11}\theta_{h00}/n_h.
\end{aligned}$$

An unbiased estimator of the variance of $\hat{\pi}$ is given by

$$\hat{V}(\hat{\pi}) = \sum_{h=1}^{L} Z_h^2 \left\{ \frac{1}{4(n_h - 1)} \left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - (2\hat{\pi}_h - 1)^2 \right] \right\}. \tag{7}$$

Now we look into the proportional and optimum allocation methods. The proportional allocation method assigns sample size n proportional to stratum size N_h. Since $n_h = n\frac{N_h}{N}$, the variance of $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \frac{1}{n} \sum_{h=1}^{L} Z_h \left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4} \right]. \tag{8}$$

The optimum allocation method assigns sample size n_h to minimize $V(\hat{\pi}_h)$ for a specified cost of taking the sample or to minimize the cost for a specified value of $V(\hat{\pi}_h)$. The simplest cost function is of the form

$$C = c_0 + \sum_{h=1}^{L} n_h c_h, \tag{9}$$

where c_0 is a fixed cost and c_h is the survey cost per unit in stratum h.

To minimize $V(\hat{\pi}_h)$ for a specified C, for stratum h the n_h can be obtained by the use of Cauchy-Schwartz inequality.

$$n_h = n \times \frac{Z_h \sqrt{\dfrac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \dfrac{(2\pi_h-1)^2}{4}} \Big/ \sqrt{c_h}}{\sum\limits_{h=1}^{L} Z_h \sqrt{\dfrac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \dfrac{(2\pi_h-1)^2}{4}} \Big/ \sqrt{c_h}}. \tag{10}$$

Finally, the minimum variance $V(\hat{\pi})$ of $\hat{\pi}$ is given by

$$\begin{aligned} V(\hat{\pi}) = \frac{1}{n} \sum_{h=1}^{L} Z_h \sqrt{\frac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \frac{(2\pi_h-1)^2}{4}} \sqrt{c_h} \\ \times \sum_{h=1}^{L} Z_h \sqrt{\frac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \frac{(2\pi_h-1)^2}{4}} \Big/ \sqrt{c_h}. \end{aligned} \tag{11}$$

Note that the optimum value of n_h depends on the known population proportion π_h of the h^{th} stratum, thus it is to be noted that optimum allocation cannot be used in real practice unless good guesses of π_h are known.

3. A stratified double estimation of a sensitive attribute by a two randomized response model

When a population is stratified but the sizes of strata are unknown, we estimate the sensitive parameter by applying stratified double sampling to Abdelfatah et al.'s model [1]. Since the size of each stratum is not known, we first classify stratum by direct questioning for stratification criterion and estimate the sensitive proportion of each stratum by using the two stage randomized response model suggested by Abdelfatah et al. [1]. We first classify the sample of size n', which was selected by a SRSWR from the population of size N with L strata into $h(h = 1, 2, \ldots, L)$ stratum of size n'_h, by asking the question, "Do you belong to stratum h?" directly. The population proportion Z_h and sample proportion z_h of group h are given respectively by

$$\begin{aligned} Z_h &= \frac{N_h}{N}, h = 1, 2, \ldots, L: \text{ The population proportion of group } h, \\ z_h &= \frac{n'_h}{n'}, h = 1, 2, \ldots, L: \text{ The sample proportion of group } h, \end{aligned}$$

where z_h is an unbiased estimator of Z_h.

The second sample is a stratified random sample of size n in which units are drawn from stratum h. Usually the second sample of size n_h in stratum h is a SRSWR from n'_h. Each respondent of $n_h (n = \sum_{h=1}^{L} n_h)$ is requested to answer the selected questions following the procedure of Abdelfatah et al.'s two stage randomized response model.

The maximum likelihood estimator $\hat{\pi}_d$ of π is given by

$$\hat{\pi}_d = \sum_{h=1}^{L} z_h \left[\frac{1}{2} + \frac{(n_{h11}/n_h - n_{h00}/n_h)B_h + (n_{h10}/n_h - n_{h01}/n_h)C_h}{2(B_h^2+C_h^2)} \right], \tag{12}$$

Theorem 3. $\hat{\pi}_d$ is an unbiased estimator of π.

Proof It follows from the fact that $E(n_{hij}/n) = \theta_{hij}$, $h = 1, 2, \ldots, L$; $i = 0, 1$; $j = 0, 1$.

$$\begin{aligned} E(\hat{\pi}_d) &= E_1\left\{E_2\left[\sum_{h=1}^{L} z_h\left(\frac{1}{2} + \frac{(n_{h11}/n_h - n_{h00}/n_h)B_h + (n_{h10}/n_h - n_{h01}/n_h)C_h}{2(B_h^2 + C_h^2)}\right)|z_h\right]\right\} \\ &= E_1\left[\sum_{h=1}^{L} z_h\left(\frac{1}{2} + \frac{(\theta_{h11} - \theta_{h00})B_h + (\theta_{h10} - \theta_{h01})C_h}{2(B_h^2 + C_h^2)}\right)\right] \\ &= \sum_{h=1}^{L} Z_h\left(\frac{1}{2} + \frac{(\theta_{h11} - \theta_{h00})B_h + (\theta_{h10} - \theta_{h01})C_h}{2(B_h^2 + C_h^2)}\right) = \sum_{h=1}^{L} Z_h\pi_h = \pi \end{aligned}$$

Theorem 4. The variance of $\hat{\pi}_d$ is given by

$$\begin{aligned} V(\hat{\pi}_d) = \frac{1}{n'}&\left\{\sum_{h=1}^{L} Z_h\left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right] + \sum_{h=1}^{L} Z_h(\pi_h - \pi)^2\right\} \\ &+ \sum_{h=1}^{L} \frac{Z_h}{n'}\left(\frac{1}{v_h} - 1\right)\left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right], \end{aligned} \tag{13}$$

where v_h is a fixed value, $0 \leqslant v_h = n_h/n'_h \leqslant 1$.

Proof Suppose that the units were measured on all n'_h just on the random subsample of n_h, the proof is easily obtained:

$$\hat{\pi}_d = \sum_{h=1}^{L} z_h\hat{\pi}_h = \sum_{h=1}^{L} z_h\hat{\pi}'_h + \sum_{h=1}^{L} z_h(\hat{\pi}_h - \hat{\pi}'_h).$$

where $\hat{\pi}'_h$ is the estimator for the sensitive attribute in the first phase with sample size n'_h.

The variance of first term on the right is

$$V\left(\sum_{h=1}^{L} z_h\hat{\pi}'_h\right) = \frac{1}{n'}\left\{\sum_{h=1}^{L} Z_h\left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right] + \sum_{h=1}^{L} Z_h(\pi_h - \pi)^2\right\}$$

and the variance of second term on the right is obtained as follows

$$E_1\left[V_2\left(\sum_{h=1}^{L} z_h(\hat{\pi}_h - \hat{\pi}'_h)\right)\right] = E_1\left\{\sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{n'_h}\right) z_h^2\left[\frac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right]\right\}.$$

Since $n_h = v_h n'_h = v_h z_h n'$,

$$\begin{aligned} E_1\left[V_2\left(\sum_{h=1}^{L} z_h(\hat{\pi}_h - \hat{\pi}'_h)\right)\right] &= E_1\left\{\sum_{h=1}^{L} \frac{z_h}{n'}\left(\frac{1}{v_h} - 1\right)\left[\frac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \frac{(2\pi_h-1)^2}{4}\right]\right\} \\ &= \sum_{h=1}^{L} \frac{Z_h}{n'}\left(\frac{1}{v_h} - 1\right)\left\{\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right\}. \end{aligned}$$

We have $V(\hat{\pi}_d)$ of Eq. (13) from the above results.

Now we look into the proportional and optimum allocation methods. Under proportional allocation, the variance of the estimator $\hat{\pi}_{d(p)}$ is given by

$$V(\hat{\pi}_{d(p)}) = \frac{1}{n'}\sum_{h=1}^{L} Z_h(\pi_h - \pi)^2 + \frac{1}{n}\sum_{h=1}^{L} Z_h\left[\frac{B_h^2(E_h + F_h) + C_h^2(G_h + H_h)}{4(B_h^2 + C_h^2)^2} - \frac{(2\pi_h - 1)^2}{4}\right] \tag{14}$$

In order to obtain the optimum values of n' and v_h, which minimize the variance $V(\hat{\pi}_d)$ for a specified cost function, we set the cost function as follows:

$$C = n'c' + \sum_{h=1}^{L} n_h c_h, \tag{15}$$

where c' is a fixed cost and c_h is the cost per unit in stratum h.

Since n_h is a random variable, we minimize the expected value of Eq. (15) to obtain the optimum values of n' and v_h:

$$E(C) = C^* = c'n' + \sum_{h=1}^{L} c_h E(n_h) = c'n' + n' \sum_{h=1}^{L} c_h v_h Z_h. \tag{16}$$

By the application of the Cauchy-Schwartz inequality to the product $C^*V(\hat{\pi}_d)$, we obtain v_h, which minimizes $C^*V(\hat{\pi}_d)$:

$$v_h = \sqrt{\frac{c'}{c_h} \frac{\dfrac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \dfrac{(2\pi_h-1)^2}{4}}{\sum\limits_{h=1}^{L} Z_h(\pi_h-\pi)^2}}. \tag{17}$$

The value of n' is obtained from the expected cost Eq. (16),

$$n' = \frac{C^*}{c' + \sum\limits_{h=1}^{L} c_h Z_h \sqrt{\dfrac{c'}{c_h} \dfrac{\dfrac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \dfrac{(2\pi_h-1)^2}{4}}{\sum\limits_{h=1}^{L} Z_h(\pi_h-\pi)^2}}} \tag{18}$$

By substitution of the optimum v_h and n', the minimum variance of $\hat{\pi}_{d(o)}$ is given by

$$V(\hat{\pi}_{d(o)}) = \frac{1}{C^*} \left[\sqrt{c'} \sqrt{\sum_{h=1}^{L} Z_h(\pi_h-\pi)^2} + \sum_{h=1}^{L} Z_h \sqrt{\frac{B_h^2(E_h+F_h)+C_h^2(G_h+H_h)}{4(B_h^2+C_h^2)^2} - \frac{(2\pi_h-1)^2}{4}} \sqrt{c_h} \right]^2. \tag{19}$$

Again note that the optimum value of n_h depends on the known population proportion π_h of the h^{th} stratum, thus it is to be noticed that optimum allocation cannot be used in real practice unless good guesses of π_h are known. In practice, the use of either proportional allocation or equal allocation is suggested while estimating proportion of a sensitive attribute using stratified randomized response sampling.

4. Efficiency comparisons

4.1. Stratified estimation vs. abdelfatah et al.'s [1] estimation

The estimator of a sensitive attribute and its variance under Abdelfatah et al.'s [1] model is given by

$$\hat{\pi}_s = \frac{1}{2} + \frac{(n_{11}/n - n_{00}/n)B + (n_{10}/n - n_{01}/n)C}{2(B^2+C^2)},$$

Table 3
The cases of RE over than 1 under $Z_1 = Z_2 = 0.5$

π	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Frequencies	0	131,220	367,416	472,392	524,880	472,392	367,416	131,220	0	2,466,936
%	0.00	5.32	14.89	19.15	21.28	19.15	14.89	5.32	0.00	100.00

Table 4
The cases of RE over than 1 under $Z_1 = 0.3, Z_2 = 0.7$

π	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Frequencies	0	144,342	328,050	472,392	524,880	472,392	328,050	144,342	0	2,414,448
%	0.00	5.98	13.59	19.57	21.74	19.57	13.59	5.98	0.00	100.00

$$V(\hat{\pi}_s) = \frac{B^2(E+F) + C^2(G+H)}{4n(B^2+C^2)^2} - \frac{(2\pi-1)^2}{4n}, \tag{20}$$

where

$$\begin{aligned}
B &= (1-W)P + (1-Q)T + W + Q - 1,\\
C &= W - Q + (1-W)P - (1-Q)T,\\
E &= WQ + W(1-Q)T + (1-W)PQ + (1-W)P(1-Q)T,\\
F &= (1-W)(1-P)(1-Q)(1-T),\\
G &= (1-Q)(1-T)[W + (1-W)P],\\
H &= (1-W)(1-P)[Q + (1-Q)T].
\end{aligned}$$

The relative efficiency (RE) of our suggested model to Abdelfatah et al.'s [1] model is the ratio $V(\hat{\pi}_s)/V(\hat{\pi})$.

$$RE = \frac{V(\hat{\pi}_s)}{V(\hat{\pi})}$$

Values of RE greater than 1 indicate our suggested model is more efficient than Abdelfatah et al.'s [1] model. In order to calculate RE empirically, we assume the population has two strata; n = 10,000, $n_1 = n_2$ = 5,000, n_1 = 3,000 and n_2 = 7,000, for each $Z_1 = Z_2 = 0.5$ and $Z_1 = 0.3$, $Z_2 = 0.7$ and we increase π, π_1 and π_2 from 0.1 to 0.9 by 0.1. For each stratum we calculate RE for the possible combinations (4,782,969) from the values of $P = P_1 = P_2, T = T_1 = T_2, W = W_1 = W_2$ and $Q = Q_1 = Q_2$ where each of parameters (P, T, W and Q) takes the values increasing from 0.1 to 0.9 by 0.1.

Under $Z_1 = Z_2 = 0.5$, the proposed estimator is more efficient than Abdelfatah et al.'s estimator in about 51.58% (2,466,936) of cases. We rearrange it according to the values of π as shown in Table 3. As shown in Table 3, the RE is symmetric around $\pi = 0.5$, which has the highest frequency.

Under $Z_1 = 0.3$, $Z_2 = 0.7$, the proposed estimator is more efficient than Abdelfatah et al.'s estimator in about 50.48% (2,414,448) of cases. We arrange it according to the values of π as shown in Table 4. As shown in Table 4, the RE is symmetric around $\pi = 0.5$, which has the highest frequency.

4.2. Stratified estimation vs. stratified double estimation

The difference between Eq. (6) with known stratum size and Eq. (13) with unknown stratum size is given by

$$\frac{1}{n'}\sum_{h=1}^{L} Z_h(\pi_h - \pi)^2. \tag{21}$$

This is a variance increment resulting from unknown stratum size.

5. Conclusions

We develop the estimation of sensitive attribute of the population which is composed of a number of strata by applying stratified sampling to Abdelfatah et al.'s model [1]. We estimate the sensitive parameter in the case of knowing the size of stratum, and check the effect of the proportional allocation method and optimum allocation method. We extend it to the case of not knowing the size of stratum, and estimate the sensitive parameter by applying stratified double sampling to Abdelfatah et al.'s model [1]. Finally, we compare the efficiency of our suggested estimator with Abdelfatah et al's existing estimator.

We have numerical comparisons with some conditions under same stratum size and different stratum size. In the former case, the proposed estimator is more efficient than Abdelfatah et al.'s estimator in 51.58% (2,466,936 out of 4,782,969) of the cases. In the latter case, the proposed estimator is more efficient than Abdelfatah et al.'s estimator in 50.48% (2,414,448 out of 4,782,969) of the cases.

Acknowledgment

The authors are thankful to the Editor-in-Chief Dr. Sarjinder Singh and a learned referee for the valuable comments to bring the original manuscript in the present form.

References

[1] S. Abdelfatah, R. Mazloum and S. Singh, Efficient use of a two-stage randomized response procedure, *Brazilian Journal of Probability And Statistics*, in press 2012.
[2] I.S. Grewal, M.L. Bansal and S.S. Sidhu, Population mean corresponding to Horvitz–Thompson's estimator for multi-characteristics using randomized response technique, *Model Assisted Statistics and Applications* **1** (2005–2006), 215–220.
[3] Z. Hong, Estimation of mean in randomized response surveys when answers are incompletely truthful, *Model Assisted Statistics and Applications* **1** (2005–2006), 221–230.
[4] M. Javed and I.S. Grewal, On the relative efficiencies of randomized response devices with Greenberg unrelated question model, *Model Assisted Statistics and Applications* **1** (2005–2006), 291–297.
[5] P.K. Mahajan, P. Sharma and R.K. Gupta, Optimum stratification for allocation proportional to strata totals for scrambled response, *Model Assisted Statistics and Applications* **2**(2) (2007), 81–88.
[6] P.K. Mahajan, Optimum stratification for scrambled response with ratio and regression methods of estimation, *Model Assisted Statistics and Applications* **1** (2005–2006), 17–22.
[7] N.S. Mangat, An improved randomized response strategy, *Journal of the Royal Statistical Society: Series B* **56** (1994), 93–95.
[8] N.S. Mangat and R. Singh, An alternative randomized response procedure, *Biometrika* **77** (1990), 439–442.
[9] O. Odumade and S. Singh, Efficient use of two decks of cards in randomized response sampling, *Communication in Statistics – Theory and Methods* **38** (2009), 439–446.
[10] P.F. Perri, Modified randomized devices for simmons' model, *Model Assisted Statistics and Applications* **3**(3) (2008), 233–239.
[11] J.-B. Ryu, J.-M. Kim, T.-Y. Heo and C.G. Park, On stratified randomized response sampling, *Model Assisted Statistics and Applications* **1** (2005–2006), 31–36.
[12] S.S. Sidhu and M.L. Bansal, Estimator of population total using Rao, Hartley and Cochran's scheme using optional randomized response technique in multi-character surveys, *Model Assisted Statistics and Applications* **3**(3) (2008), 259–267.
[13] S.L. Warner, Randomized response, a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60** (1965), 63–69.
[14] Z. Yan, J. Wang, J. Lai and W. Hua, Ratio imputation method for handling item-nonresponse in Eichhorn model, *Model Assisted Statistics and Applications* **3**(2) (2008), 89–98.